Tracking Visual Memes in Rich-Media Social Communities

Lexing Xie[*], **Apostol Natsev**[†], **John R Kender**[‡], **Matthew Hill**[†], **John R Smith**[†]
[*]The Australian National University, Canberra, Australia
[†]IBM T J Watson Research Center, Hawthorne, NY, USA
[‡]Columbia University, New York, NY

Abstract

We propose tools and methods to track visual memes on community-centric rich-media repositories, such as YouTube. Visual memes refer to frequently reposted short video segments. Our method can be used to monitor the reflections of real-world events in rich media, including images and videos. We first design a large-scale event-based social video collection system to continuously monitor events that unfold in real-time. We design a scalable detection algorithms that can detect visual memes with over 96% precision and 80% recall. Visual memes are used for various analysis such as tracking the fraction of original content, extracting the iconic picture of an event, inferring influential users in the community, and so on. We present example observations on several real-world video collections from YouTube, containing up to 1.2 million video shots, including a compact taxonomy of authors into "traditional news media", "citizen buzz leaders", and "mavens".

1 Introduction

Media-rich online communities, notably YouTube, have become virtual worldwide bazaars for videos of almost every type. With more than 24 hours of video being added every minute, YouTube is a living marketplace of ideas and a vibrant recorder of current events. The ease of publishing and sharing videos has outpaced the progress of modern search engines—leaving users to see only a fraction of their subject. This information overload problem is particularly prominent for linear media (such as audio, video, animations), where at-a-glance impressions are hard to develop and are often unreliable. The goal of this research is to develop a tool to reliably track online video propagation, in a manner similar to *hashtags* and *retweets* that are prevalent in the text-based micro-blogging community. Such a system will be useful in many different domains, such as brand and image monitoring, journalistic content selection, or better social data sampling and storage systems.

We propose methods and tools to track visual memes for making sense of video buzz. A *meme* is defined as[1] a cultural unit (e.g., an idea, value, or pattern of behavior) that is passed from one person to another in social settings. We define a *visual meme* as a short segment of video that is frequently remixed and reposted by more than one uploader. Example visual memes are shown in Figures 1 and 2, represented in a static keyframe format. We can see that despite many variations in the videos that contain them (such as size, coloring, captions, editing), each meme instance is semantically consistent.

Our work on tracking video remix mainly relates to two topics in existing research. The first topic is on quoting, duplication, and reposting in online information networks. One well-known example is the use of the RT (re-tweet) tag on Twitter (Kwak et al. 2010; Cha et al. 2010), where users often quote the original message verbatim, having little freedom for remixing and context changes within the 140 character limit. Another example is MemeTracker (Leskovec, Backstrom, and Kleinberg 2009), which tracks the lifecycles of popular phrases among blogs and news websites. Prior studies have shown that the frequency of video reuse can be used as an implicit video quality indicator (Schmitz et al. 2006). The second topic is on detecting near-duplicates in images and video collections (Hampapur, Hyun, and Bolle 2002). Current detection methods perform matching based on image sequence, frame, or local image points (Tan et al. 2008), and can be deployed on large distributed systems (Liu, Rosenberg, and Rowley 2007). Compared to prior work, this paper is the first to propose visual memes as the unit of social information diffusion, using a scalable method to track them.

We design and implement a large-scale event-based social video monitoring and visual content analysis system using targeted keyword querying. We propose scalable meme-tracking algorithms to extract all memes from over a million video shots, using robust visual matching and approximate nearest neighbor indexing techniques. Visual memes are then used as a tool to characterize on-topic video content. We present several observations on a large-scale event datasets around the Iranian election in 2009. We found that over half of the videos in an event collection are not original; that video view count is a poor proxy for the likelihood of a video being reposted; and that the roles of traditional news media and citizen journalists can be characterized using meme-related measures that capture diffusion influence.

2 Memes and Video Content Duplication

Visual memes are defined as frequently reposted video segments or images. Video-making requires significant effort and time, so we regard the reposting of a video meme to

[1]http://wordnetweb.princeton.edu/perl/webwn?s=meme

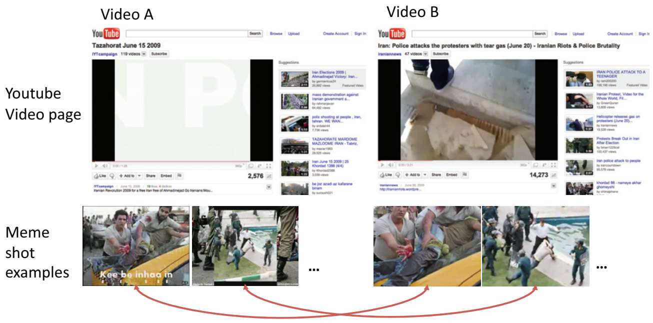

Figure 1: Two YouTube videos that share multiple different memes. Note that it is impossible to tell from metadata that they shared content, and that the appearance of the remixed shots (bottom row) has large variations.

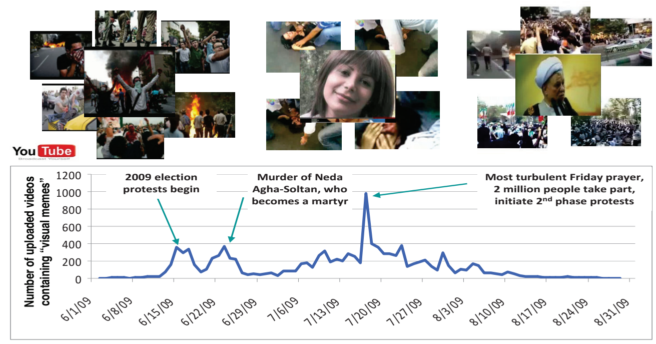

Figure 2: Meme videos on the Iran3 topic: representative visual memes on a timeline, June-August 2009.

be a much stronger endorsement requiring much more effort than simply viewing, commenting on, or linking to the content. In this paper, we will use "meme" in two ways: singularly, to refer to individual video segment instances, which are visualized as representative icons (as in Figure 2), and collectively, to refer to the entire equivalence class of re-posted near-duplicate video segments, which are visualized as sets of similar keyframes (such as the linked frames in Figure 1). This dual use reflects the observation that users tend not to repeat others' entire videos unaltered, but neither do they often create highly edited original stories. The unit of interaction appears to be video segments, consisting of one or a few contiguous shots, with only minor modifications such as video formatting changes (aspect ratio, color, contrast, gamma), and video production edits (the superimposition of text, captions, borders, transition effects).

A visual meme is a basic unit of information that propagates through the video information network relatively intact. Compared to a user tag or hashtag, it is a representation of a specific idea rather than an overall topic. Compared to a re-tweet that repeats a short message verbatim, it carries forward part of the idea and often remixes with new ones. Its closest analogy in textual media is the meme phrase (Leskovec, Backstrom, and Kleinberg 2009).

3 Monitoring Events on YouTube

It is impossible to externally monitor all YouTube content. Instead, we use a few generic, time-insensitive text queries as pre-filters to narrow down the content scope. The queries are manually designed to capture a generic topic theme, as well as any generally understood cause, phenomena, and consequences of the topic. For example, our query on the "global warming" topic consists of *global warming, climate change, green house gas, CO_2 emission*, and so on. We have aimed to create queries covering the main invariant aspects of a topic, but automatic time-varying query expansion is open for future work. We use the YouTube API to query and extract as many available videos for each query that the API will allow. We filter the results to remove videos that have responded to multiple queries, or those whose YouTube identifier match one that had previously been gathered. Then, for each unique video, we segment shots, extract keyframes, and extract visual features from each keyframe. We also extract information from the associated metadata, such as author, publish date, view counts, and free-text title and descriptions.

Figure 2 illustrates the volume of meme videos (unique videos containing one or more memes) for the Iranian Politics topic (dataset Iran3 in Table 1). There are hundreds to thousands per day, with three prominent peaks in June-August 2009 corresponding to important events in the real world[2]. The first mid-June peak reflects a highly controversial election prompting massive protests and violent clashes. A second mid-June peak captures the virality of an amateur video on the shooting of Neda Soltan, which became the symbol for the whole event. A third peak in mid-July corresponds to a Friday prayer sermon that drew over 2 million people, an event described as "the most critical and turbulent Friday prayer in the history of contemporary Iran"[2].

4 Scalable Visual Meme Detection

There are two main challenges in detecting visual memes in a large collection. The first is the noise in matching video segments, since remixed online video segments tend to vary widely in appearance (Section 2). The second is the overall computational complexity of the matching. Finding all pairs of near-duplicates by matching all N shots against each other has a complexity of $O(N^2)$, which is infeasible for a typical collection containing millions of video shots.

Our solution to the first challenge is robust keyframe matching, where a keyframe is representative of a video shot, segmented using temporal feature differences. We preprocess the frame by removing trivial (e.g. blank) matches. we then extract the *color correlogram* (Huang et al. 1999) feature for each frame to capture the local spatial correlation of pairs of colors. The color correlogram is rotation-, scale-, and to some extent, viewpoint-invariant. We also use a "cross"-layout that extracts the descriptor only from horizontal and vertical central image stripes, thereby emphasizing the center portion of the image and disregard-

[2]See timeline: `http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests`

ing the corners. This layout improves robustness with respect to text and logo overlay, borders, crops, and shifts. The keyframe matching uses a query-adaptive threshold to normalize among the query frame, and among the different feature dimensions. This threshold is tuned on a training set.

Our solution to the complexity challenge is to use an indexing scheme for fast approximate nearest neighbor (ANN) look-up. We use the FLANN Library (Muja and Lowe 2009) to automatically select the best indexing structure and its appropriate parameters for a given dataset. Our frame features have over 300 dimensions, and we empirically found that setting the number of nearest-neighbor candidate nodes m to $\sqrt{N}$ can approximate k-NN results with approximately 0.95 precision. In running in $O(N\sqrt{N})$ time, it achieves two to three decimal orders of magnitude speed-up over exact nearest neighbor search. Furthermore, each FLANN query results in an incomplete set of near-duplicate pairs, we perform transitive closure on the neighbor relationship to find equivalence classes of near-duplicate sets. We use an efficient set union-find algorithm that runs in amortized time of $O(E)$, where E is the number of matched pairs (Galler and Fisher 1964), which is again $O(N\sqrt{N})$.

5 Using Memes to Estimate Influence

One application for visual memes is in estimating the impact of content and of authors within a content sharing network. Visual memes can be viewed as *links* between videos and their creators that share the same unit of visual expression. We derive a link-based measure, called *diffusion influence index*, to depict the influence of a meme and its author.

Denote a video (or multimedia document) as d_i in event collection $\mathcal{D}$, with $i = 1, \ldots, N$. Each video is authored (i.e., uploaded) by author $a(d_i)$ at time $t(d_i)$, with $a(d_i)$ taking its value from a set of authors $\mathcal{A} = \{a_r, r = 1, \ldots, R\}$. Each video document d_i contains a collection of memes, $\{m_1, m_2, \ldots, m_{K_i}\}$ from a meme dictionary $\mathcal{M}$. We compute the in-degree (and out-degree) of each meme m in video d_i, as the number of other videos containing meme m that appeared before (and after) d_i.

$$\begin{aligned} \zeta^{in}_{i,m} &= \Sigma_j I\{m \in d_j,\ t(d_j) < t(d_i)\} \\ \zeta^{out}_{i,m} &= \Sigma_j I\{m \in d_j,\ t(d_j) > t(d_i)\} \end{aligned} \tag{1}$$

where $I\{\cdot\}$ is the indicator function that takes a value of 1 when its argument is true, and 0 otherwise. Intuitively, $\zeta^{in}_{i,m}$ and $\zeta^{out}_{i,m}$ captures the number of videos are potential sources and potential followers for meme m in d_i. The video influence index χ_i is defined for each video document d_i as the ratio of its out-degree over its in-degree, aggregated over all memes (Equation 2). The smoothing constant in the denominator accounts for d_i itself. The total author influence index $\hat{\chi}_r$ is the aggregate χ_i over all videos from author a_r (Equation 3).

$$\chi_i = \Sigma_m \frac{\zeta^{out}_{i,m}}{1 + \zeta^{in}_{i,m}} \tag{2}$$

$$\hat{\chi}_r = \Sigma_{\{i, a(d_i) = a_r\}}\ \chi_i \tag{3}$$

Intuitively, this influence index gives a higher score to a video containing a very popular meme that was uploaded early.

6 Experiments and Observations

We present example observations on three test datasets, collected from YouTube using the targeted-querying and collection procedures described in Section 3. Table 1 contains an overview of the datasets. The Iran3 set is about Iranian domestic politics and related international events during the 3-month period of summer 2009. The Iran1 set is a subset of Iran3 that was collected during the first of the 3 months; most of its videos are about the election in mid-June and its associated political outbreaks. The Housing set is about the housing market in the 2008-09 economic crisis; this hand-annotated set was used as a validation set for tuning the visual meme detection algorithms. We perform visual meme detection as described in Section 4. We additionally filter the meme clusters identified by the detection system, by removing singletons belonging to a single video or a single author.

topic	#videos	#authors	#shots	upload time
Iran3	23,049	4,681	1,255,062	08/07∼08/09
Iran1	5,429	2,393	210,259	09/07∼07/09
Housing	2,446	654	71,872	08/07∼08/09

Table 1: Summary of YouTube event data sets.

Meme detection performance

On the Iran3 set of over 1 million shots, feature extraction takes around 7 hours on a quad-core CPU, and indexing and querying with FLANN takes 5 to 6 CPU hours. We measure the meme detection performance using ground truth from the Housing dataset, which contains $\sim 15,000$ near-duplicate keyframe pairs and $\sim 25,000$ non-duplicate keyframe pairs, sampled from multiple runs of k-means clustering and nearest-neighbor search. Each pair was manually verified, and the collection specifically includes many pairs near the decision boundary. We compute the near-duplicate equivalence classes as described in Section 4, and calculate precision (P) and recall (R) on the labeled pairs. The results are shown on Figure 3 left for varying values of the threshold parameter τ. We note that the performance is generally quite high with $P > 95\%$. For the rest of our analysis, we use an operating point with high recall $P = 96.6\%,\ R = 80.7\%$.

Content views and re-posting probability

In our video collections, the behavior of remixing and re-posting is quite dominant – over 58% of the videos and 70% of the authors contain visual memes for Iran3. However, we observe that video popularity is a poor indicator of how likely it is to be re-posted. In the Iran3 set of more than 23K videos, for example, the four most popular videos have no memes and have nothing to do with the topic, and likewise for 7 of the first 10. One has to get beyond the first 1,600

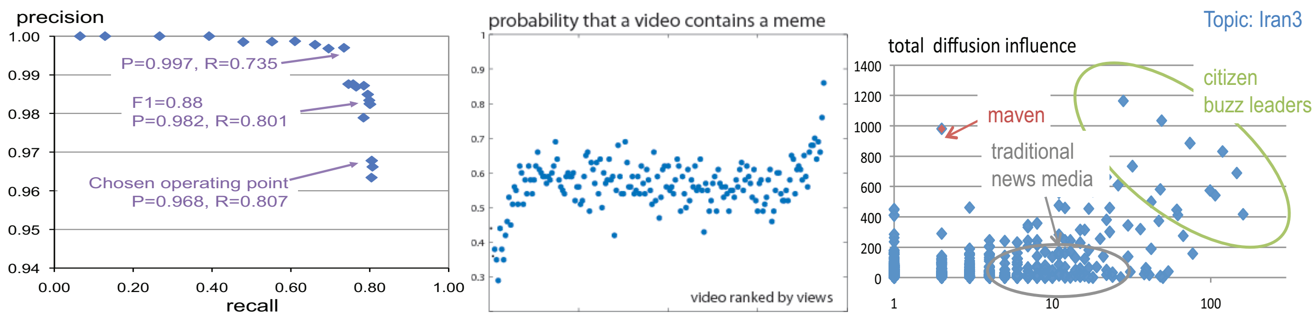

Figure 3: From left to right: Performance of visual meme detection method on the Housing dataset. Video views vs. meme probability on Iran3 set.

most popular videos before the likelihood of having near-duplicates passes the average for the dataset, at about 0.58 (see Figure 3 middle). There are several reasons for this mismatch. Among the video entries returned by YouTube search API, the most viewed are often not related to the topic – the one with the highest view-count is a music video irrelevant to Iranian politics. Such videos also tend to be part of a production (e.g. promotion for a song), which bears lesser value for re-posting and re-interpretation. Moreover, it is influenced by a "rich-get-richer" effect due to content recommendations and promotions on YouTube site. In short, popularity is a poor proxy for relevance or importance.

Diffusion index for content and authors

We compute the diffusion index for authors according to Equation 3 on dataset Iran3. On the right of Figure 3 we plot the total diffusion influence $\hat{\chi}_r$ versus the number of videos produced by each author. We can see a few distinct types of contributors. We call one type "maven" (marked in red), who post only a few videos that tend to be massively remixed and reposted – this particular maven was among the first to post the murder of Neda Soltan, and one other instance of student murder on the street. The former become the icon of the entire event timeline. A second group can be dubbed "citizen buzz leaders" (circled in green), who tend to produce a large number of videos with high total diffusion factor, yet relatively low influence per video. They aggregate notable content and come relatively late in the timeline, which is penalized by the influence factor. We examined the YouTube channel pages for a few authors in this group, and they seem to be voluntary political activists with screen-names like "iranlover100". Some of their videos are slide shows of iconic images. Note that traditional news media, such as AljezeeraEnglish, AssociatedPress and so on (circled in gray) are ranked rather low for this topic, partially because the Iran government severely limited international media participation in the event; most of the event buzz was driven by citizens.

7 Conclusions

In this paper, we proposed using visual memes for tracking and monitoring of social diffusion on YouTube. We described methods and tools for large-scale content monitoring, and a scalable algorithm for extracting visual memes with over 98% precision and 80% recall. We have observed in real-world youtube collection containing over 1 million shots, that over half of the videos in typical event collections contain re-mixed content; video view counts are poorly correlated with the probability of being remixed and reposted. Visual memes can be used to characterize diffusion influence, such as capturing the roles that citizen journalists and traditional news media play. Future work includes exploring the meanings and evolution of memes, or understanding and predicting visual meme virality.

References

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*.

Galler, B. A., and Fisher, M. J. 1964. An improved equivalence algorithm. *Commun. ACM* 7(5):301–303.

Hampapur, A.; Hyun, K.; and Bolle, R. 2002. Comparison of sequence matching techniques for video copy detection. In *Conf. on Storage and Retrieval for Media Databases*.

Huang, J.; Kumar, S.; Mitra, M.; Zhu, W.; and Zabih, R. 1999. Spatial color indexing and applications. *International Journal of Computer Vision* 35(3).

Kwak, H.; Lee, C.; Park, H.; ; and Moon, S. 2010. What is Twitter, a Social Network or a News Media? . In *WWW*.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proc. KDD*.

Liu, T.; Rosenberg, C.; and Rowley, H. A. 2007. Clustering billions of images with large scale nearest neighbor search. In *IEEE Workshop on Applications of Computer Vision*, 28.

Muja, M., and Lowe, D. G. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In *Intl. Conf. on Computer Vision Theory and Applications*.

Schmitz, P.; Shafton, P.; Shaw, R.; Tripodi, S.; Williams, B.; and Yang, J. 2006. International remix: video editing for the web. In *Proc. ACM Multimedia*, 798. ACM.

Tan, H.-K.; Wu, X.; Ngo, C.-W.; and Zhao, W.-L. 2008. Accelerating near-duplicate video matching by combining visual similarity and alignment distortion. In *Proc. ACM Multimedia*, 861.